Int J Comput Vis (2018) 126:440–456
https://doi.org/10.1007/s11263-017-0997-7

Deep Multimodal Fusion: A Hybrid Approach

Mohamed R. Amer[1] · **Timothy Shields**[1] · **Behjat Siddiquie**[1] · **Amir Tamrakar**[1] · **Ajay Divakaran**[1] · **Sek Chai**[1]

Received: 16 February 2016 / Accepted: 6 February 2017 / Published online: 20 February 2017

Abstract We propose a novel hybrid model that exploits the strength of discriminative classifiers along with the representation power of generative models. Our focus is on detecting multimodal events in time varying sequences as well as generating missing data in any of the modalities. Discriminative classifiers have been shown to achieve higher performances than the corresponding generative likelihood-based classifiers. On the other hand, generative models learn a rich informative space which allows for data generation and joint feature representation that discriminative models lack. We propose a new model that jointly optimizes the representation space using a hybrid energy function. We employ a Restricted Boltzmann Machines (RBMs) based model to learn a shared representation across multiple modalities with time varying data. The Conditional RBMs (CRBMs) is an extension of the RBM model that takes into account short term temporal phenomena. The hybrid model involves augmenting CRBMs with a discriminative component for classification. For these purposes we propose a novel Multimodal Discriminative CRBMs (MMDCRBMs) model. First, we train the MMDCRBMs model using labeled data by training each modality, followed by training a fusion layer. Second, we exploit the generative capability of MMDCRBMs to activate the trained model so as to generate the lower-level data corresponding to the specific label that closely matches the actual input data. We evaluate our approach on ChaLearn dataset, audio-mocap, as well as the Tower Game dataset, mocap-mocap as well as three multimodal toy datasets. We report classification accuracy, generation accuracy, and localization accuracy and demonstrate its superiority compared to the state-of-the-art methods.

Keywords Deep learning · Conditional Restricted Boltzmann Machines · Hybrid · Generative · Discriminative · Multimodal fusion · Gesture recognition · Social interaction modeling

Communicated by Cordelia Schmid ,V. Lepetit.

Mohamed R. Amer and Timothy Shields have contributed equally to this work

✉ Mohamed R. Amer
mohamed.amer@sri.com

Timothy Shields
Timothy.Shields@sri.com

Behjat Siddiquie
Behjat.Siddiquie@sri.com

Amir Tamrakar
Amir.Tamrakar@sri.com

Ajay Divakaran
Ajay.Divakaran@sri.com

Sek Chai
Sek.Chai@sri.com

[1] SRI International, Princeton, NJ 08540, USA

1 Introduction

Many real life events are inherently multimodal with each modality containing information useful for detecting or recognizing the event. Despite this, there is a significant amount of work focused on modeling and recognizing events using a single modality, neglecting other sources of information when available. While this might be sufficient for certain problems, it is inadequate when the events to be detected are complex and subtle. Humans are capable of combining cues from multiple modalities to reason about specific events. Therefore when multiple, information rich, modalities are present, it becomes important to jointly interpret

and reason about the information from each modality. While jointly modeling multiple modalities, the temporal information within and across modalities also needs to be accounted for. Following the human cognitive system, we propose to solve the multimodal fusion using a biologically-inspired model namely Conditional Restricted Boltzmann Machines (CRBMs) (Taylor et al. 2011).

Discriminative models focus on maximizing the separation between classes, however, they are often uninterpretable or require some heavy reverse engineering (Zeiler and Fergus 2014). On the other hand, generative models focus solely on modeling distributions and are often unable to incorporate higher level knowledge. Hybrid models tend to address these problems by combining the advantages of discriminative and generative models. They encode higher level knowledge as well as model the distribution from a discriminative perspective. We propose a novel hybrid model that allows us to recognize classes, correlate features, and generate input data.

The CRBMs are non-linear generative models for modeling time series data. They use an undirected model with binary latent variables connected to a number of visible variables. A CRBM based generative model enables modeling short-term multimodal phenomenon and also allows us to deal with missing data by generating it within or across modalities. We propose a hybrid model to acquire the benefits of a discriminative classifier. The hybrid model involves enhancing the CRBM with a discriminative component based on the work of (Larochelle and Bengio 2008). Leading to a superior classification performance, while also allowing us to model temporal dynamics. We evaluate our approach on multiple audio-visual datasets and show how our results are comparable/superior to the state-of-the-art approaches.

Our Contributions We extend our initial work (Amer et al. 2014) and propose a new general hybrid model. Compared to Amer et al. (2014) we contributed the following:

- We propose a new jointly trained hybrid model that combines the advantages of temporal generative and discriminative models forming an extendable formal multimodal fusion framework for classifying multimodal data.
- We evaluate our model on realistic datasets and we are the first to report generation accuracy on both datasets since other work used only discriminative models.
- We extensively evaluate the effect of additional training data and the effect of additional model parameters affect our performance.

Paper Organization In Sect. 2 we discuss prior work. In Sect. 3 we give a brief background of similar models that motivate our approach, followed by a description of our hybrid model. In Sect. 4 we describe the inference algorithm. In Sect. 5 we specify our learning algorithm. In Sect. 6 we show quantitative results of our approach, followed by the conclusion in Sect. 7.

2 Prior Work

In this section we first review literature on multimodal fusion; second we review hybrid models; finally we review temporal, energy-based, deep learning.

Multimodal Fusion Deep networks have been used for multimodal fusion (Srivastava and Salakhutdinov 2012) for tags and image fusion (Ngiam et al. 2011) for audio spectrograms and image fusion. These models are generative, however, they operate on static data. Other work focused on temporal datasets (Audio-Video) such as AVEC dataset (Glodek et al. 2011). Representative work on AVEC includes generative models such as the Hidden Markov Models (HMMs) based methods (Glodek et al. 2011) and discriminative models such as the Conditional Random Fields (CRFs) (Ramirez et al. 2011). Each of these approaches separately lack the advantages of the other, whereas hybrid models, include the abilities to learn a joint representation combining the benefits of both discriminative and generative models benefits. Recently a new challenging temporal dataset, ChaLearn (Escalera et al. 2014), was released for evaluating multimodal gesture recognition. The most successful algorithm in the challenge was a discriminative deep learning Convolutional Neural Networks (CNNs) model (Neverova et al. 2014). While this approach achieves the best results (even compared to our approach), they ignore the temporal aspect of the data and modeled each modality specifically with so much engineering. Our experiments showed that jointly modeling modalities using a generative deep learning architecture helps in substantially improving both classification over (Amer et al. 2014), achieving relatively comparable results to Neverova et al. (2014) with no engineering of the architecture and generation performance that was not explored by their approach. In this paper, we focus on the modeling of multimodal data using a hybrid model.

Hybrid Models These models consist of a generative component, which usually learns a feature representation given low-level input, and a discriminative component for higher level reasoning. Recent work has empirically shown that generative models which learn a rich feature representation tend to outperform discriminative models that rely solely on handcrafted features (Perina et al. 2012). Hybrid models can be divided into three groups, joint methods (Larochelle and Bengio 2008; Druck and McCallum 2010), iterative methods (Sminchisescu et al. 2006; Fujino et al. 2008), and staged methods (Li et al. 2011; Ranzato et al. 2011; Perina et al.

2012). Joint methods optimize a single objective function which consists of both the generative and discriminative components used to learn a joint representation. They are usually learned using methods such as variational learning. Iterative methods, similar to joint methods. learn a shared representation layer using an iterative learning approach, such as Expectation Maximization, where the representations are updated using updates from the discriminative component and the generative component. Staged methods, are different than joint and iterative since both generative and discriminative components are trained separately in a staged manner. Generative representations are learned in an unsupervised manner, followed by discriminative components learned with supervision using the generative representations as a new input.

Representation Learning Deep learning have been successfully applied to many problems (Bengio 2009). Restricted Boltzmann Machines (RBMs) form the building blocks in energy based deep networks (Hinton et al. 2006; Salakhutdinov and Hinton 2006). In Hinton et al. (2006), Salakhutdinov and Hinton (2006), the networks are trained using the contrastive divergence (CD) algorithm (Hinton 2002), which demonstrated the ability of deep networks to capture the distributions over the features efficiently and to learn complex representations. RBMs can be stacked together to form deeper networks known as Deep Boltzmann Machines (DBMs), which capture more complex representations. Recently, temporal models based on deep networks have been proposed, capable of modeling a more temporally rich set of problems. These include Conditional RBMs (CRBMs) (Taylor et al. 2011) and Temporal RBMs (TRBMs) (Sutskever and Hinton 2007; Sutskever et al. 2008; Hausler and Susemihl 2012). CRBMs have been successfully used in both visual and audio domains. They have been used for modeling human motion (Taylor et al. 2011), tracking 3D human pose (Taylor et al. 2010) and phone recognition (Mohamed and Hinton 2009). TRBMs have been applied for transferring 2D and 3D point clouds (Zeiler et al. 2011), transition based dependency parsing (Garg and Henderson 2011), and polyphonic music generation (Lewandowski et al. 2012).

3 Model

Using a hybrid model allows us to take advantage of the benefits of generative models, which include filling in missing data and the benefits of a discriminative model, leading to a stronger classifier compared to purely generative models.

Rather than immediately defining our MMDCRBM model, we discuss a sequence of models, gradually increasing in complexity, so that the different components of our hybrid model can be understood in isolation. We start with the basic RBM model (Sect. 3.1), then we extend the RBM to the temporal CRBM model (Sect. 3.2), then we extend the CRBM to the multimodal MMCRBM model (Sect. 3.3). Then we make each of those three models discriminative: DRBM (Sect. 3.4), DCRBM (Sect. 3.5), and finally MMDCRBM (Sect. 3.6).

3.1 Restricted Boltzmann Machines (RBMs)

RBMs (Salakhutdinov and Hinton 2006), shown in Fig. 1a, defines a probability distribution p_{R} as a Gibbs distribution (1), where $\mathbf{v}$ is a vector of visible nodes, $\mathbf{h}$ is a vector of hidden nodes, E_{R} is the energy function, and Z is the partition function. The parameters $\boldsymbol{\theta}$ to be learned are $\mathbf{a}$ and $\mathbf{b}$ the biases for $\mathbf{v}$ and $\mathbf{h}$ respectively and the weights W. The RBM is fully connected between layers, with no lateral connections. This architecture implies that $\mathbf{v}$ and $\mathbf{h}$ are factorial given one of the two vectors. This allows for the exact computation of $p_{\mathrm{R}}(\mathbf{v}|\mathbf{h})$ and $p_{\mathrm{R}}(\mathbf{h}|\mathbf{v})$.

$$\begin{aligned} p_{\mathrm{R}}(\mathbf{v},\mathbf{h}) &= \tfrac{\exp[-E_{\mathrm{R}}(\mathbf{v},\mathbf{h})]}{Z(\boldsymbol{\theta})}, \\ Z(\boldsymbol{\theta}) &= \textstyle\sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\mathrm{R}}(\mathbf{v},\mathbf{h})], \\ \boldsymbol{\theta} &= \begin{bmatrix} \{\mathbf{a},\mathbf{b}\} & \text{-bias,} \\ \{W\} & \text{-fully connected.} \end{bmatrix} \end{aligned} \tag{1}$$

In case of binary valued data v_i is defined as a logistic function. In case of real valued data, v_i is defined as a multivariate Gaussian distribution with a unit covariance. A binary valued hidden layer h_j is defined as a logistic function such that the hidden layer becomes sparse (Taylor et al. 2011; Sutskever and Hinton 2007). The probability distributions over v, and h is defined as in (2).

$$\begin{aligned} p_{\mathrm{R}}(v_i = 1|\mathbf{h}) &= \sigma(a_i + \textstyle\sum_j h_j w_{ij}), && \text{Binary,} \\ p_{\mathrm{R}}(v_i|\mathbf{h}) &= \mathcal{N}(a_i + \textstyle\sum_j h_j w_{ij}, 1), && \text{Real,} \\ p_{\mathrm{R}}(h_j = 1|\mathbf{v}) &= \sigma(b_j + \textstyle\sum_i v_i w_{ij}), && \text{Binary.} \end{aligned} \tag{2}$$

The energy function E_{R} for binary v_i is defined as in (3).

$$E_{\mathrm{R}}(\mathbf{v},\mathbf{h}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i w_{ij} h_j, \tag{3}$$

while, the energy function E_{R} is slightly modified to allow for the real valued $\mathbf{v}$ as shown in (4).

$$E_{\mathrm{R}}(\mathbf{v},\mathbf{h}) = -\sum_i \frac{(a_i - v_i)^2}{2} - \sum_j b_j h_j - \sum_{i,j} v_i w_{ij} h_j \tag{4}$$

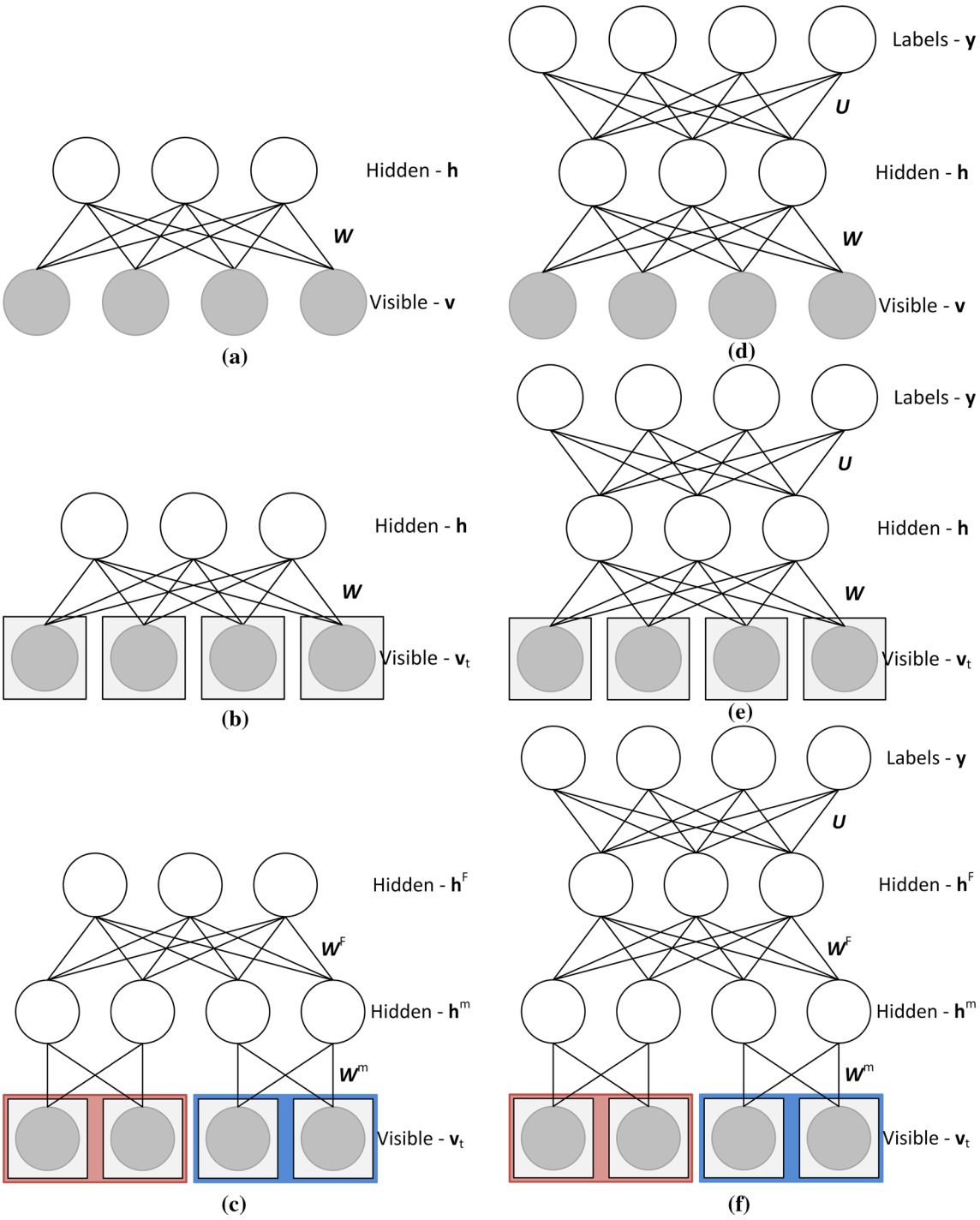

Fig. 1 This figure illustrates the progression of models described in Sect. 3. **a** RBM **b** CRBM and **c** MMCRBM are generative models that can be trained in an unsupervised manner. **d** DRBM **e** DCRBM and **f** MMDCRBM are the discriminative counter part that are trained in a supervised manner. The extension from the left column to the right column lies in adding a discriminative component (Larochelle and Bengio 2008) to the generative models. The extension across the rows is a progression from static models, to dynamic models, to multimodal dynamic models

3.2 Conditional Restricted Boltzmann Machines (CRBMs)

CRBMs (Taylor et al. 2011), shown in Fig. 1b, are a natural extension of RBMs for modeling short term temporal dependencies. A CRBM (Fig. 1) is an RBM which takes into account history from the previous time instances $t-N,\ldots,t-1$ at time t. This is done by treating the previous time instances as additional inputs. Doing so does not complicate inference. Some approximations have been made to facilitate efficient training and inference, more details are available in Taylor et al. (2011). A CRBM defines a probability distribution p_{C} as a Gibbs distribution (5).

$$\begin{aligned} &p_{\mathrm{C}}(\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t}) = \tfrac{\exp[-E_{\mathrm{C}}(\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t})]}{Z(\boldsymbol{\theta})},\\ &Z(\boldsymbol{\theta}) = \textstyle\sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\mathrm{C}}(\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t})]\\ &\boldsymbol{\theta} = \begin{bmatrix} \{\mathbf{a},\mathbf{b}\} & \text{-bias,}\\ \{A,B\} & \text{-auto regressive,}\\ \{W\} & \text{-fully connected.}\end{bmatrix} \end{aligned} \tag{5}$$

The visible vectors from the previous N time instances, denoted as $\mathbf{v}_{<t}$, influence the current visible and hidden vectors. The probability distributions are defined in (6).

$$\begin{aligned} &p_{\mathrm{C}}(v_i|\mathbf{h},\mathbf{v}_{<t}) = \mathcal{N}(c_i + \textstyle\sum_j h_j w_{ij}, 1),\\ &p_{\mathrm{C}}(h_j = 1|\mathbf{v},\mathbf{v}_{<t}) = \sigma(d_j + \textstyle\sum_i v_i w_{ij}). \end{aligned} \tag{6}$$

where,

$$\begin{aligned} &c_i = a_i + \textstyle\sum_p A_{pi} v_{p,<t},\\ &d_j = b_j + \textstyle\sum_p B_{pj} v_{p,<t}. \end{aligned} \tag{7}$$

The new energy function $E_{\mathrm{C}}(\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t})$ in (8) is defined in a manner similar to that of the RBM (4).

$$\begin{aligned} E_{\mathrm{C}}(\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t}) = &-\textstyle\sum_i (c_i - v_{i,t})^2/2 - \sum_j d_j h_{j,t}\\ &-\textstyle\sum_{i,j} v_{i,t} w_{ij} h_{j,t}, \end{aligned} \tag{8}$$

Note that A and B are matrices of concatenated vectors of previous time instances of $\mathbf{a}$ and $\mathbf{b}$.

3.3 Multimodal Conditional Restricted Boltzmann Machines (MMCRBMs)

The extension of the CRBM to multimodal is straightforward as shown in Fig. 1c. We define a Gibbs distribution over a multimodal network of stacked CRBMs, letting $p_{\mathrm{MC}}(\mathbf{v}_t^{1:M},\mathbf{h}_t^{1:M},\mathbf{h}_t^F|\mathbf{v}_{<t}^{1:M})$ denote the distribution (9). This is similar to the approach proposed in Srivastava and Salakhutdinov (2012) and Ngiam et al. (2011) except that we use CRBMs as our main building block instead of RBMs or auto-encoders. This enables us to model the temporal nature of the time-series data.

$$\begin{aligned} &p_{\mathrm{MC}}(\mathbf{v}_t^{1:M},\mathbf{h}_t^{1:M},\mathbf{h}_t^F|\mathbf{v}_{<t}^{1:M}) = \tfrac{\exp[-E_{\mathrm{MC}}(\mathbf{v}_t^{1:M},\mathbf{h}_t^{1:M},\mathbf{h}_t^F|\mathbf{v}_{<t}^{1:M})]}{Z(\boldsymbol{\theta})},\\ &Z(\boldsymbol{\theta}) = \textstyle\sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\mathrm{MC}}(\mathbf{v}_t^{1:M},\mathbf{h}_t^{1:M},\mathbf{h}_t^F|\mathbf{v}_{<t}^{1:M})],\\ &\boldsymbol{\theta} = \begin{bmatrix} \{\mathbf{a}^{1:M},\mathbf{b}^{1:M},\mathbf{e}\} & \text{-bias,}\\ \{A^{1:M},B^{1:M},C^{I:M}\} & \text{-auto regressive,}\\ \{W^{1:M},W^F\} & \text{-fully connected.}\end{bmatrix} \end{aligned} \tag{9}$$

The probability distributions are defined in (10).

$$\begin{aligned} &p_{\mathrm{MC}}(v_{i,t}^m|\mathbf{h}_t^m,\mathbf{v}_{<t}^m) = \mathcal{N}(c_i^m + \textstyle\sum_j h_j^m w_{ij}^m, 1),\\ &p_{\mathrm{MC}}(h_{j,t}^m = 1|\mathbf{h}_t^F,\mathbf{v}_t^m,\mathbf{v}_{<t}^m) = \sigma(d_j^m\\ &\quad + \textstyle\sum_k h_k^F w_{jk}^F + \sum_i v_{i,t}^m w_{ij}^m),\\ &p_{\mathrm{MC}}(h_t^F|\mathbf{h}_t^{1:M},\mathbf{h}_{<t}^{1:M}) = \sigma(f_k + \textstyle\sum_{m,j} h_t^m w_{jk}^F). \end{aligned} \tag{10}$$

where,

$$\begin{aligned} &c_i^m = a_i^m + \textstyle\sum_p A_{pi}^m v_{p,<t}^m,\\ &d_j^m = b_j^m + \textstyle\sum_p B_{pj}^m v_{p,<t},\\ &f_k = e_k + \textstyle\sum_{m,r} C_{rk}^m h_{k,<t}^m. \end{aligned} \tag{11}$$

For a multimodal CRBM, we define the joint representation (fusion) layer to be the top layer. The multimodal energy $E_{\mathrm{MC}}(\mathbf{v}_t^{1:M},\mathbf{h}_t^{1:M},\mathbf{h}_t^F|\mathbf{v}_{<t}^{1:M})$ is decomposed into two parts as shown in (12). The first part is the fusion energy for the joint representation, where $\mathbf{h}_t^F$ is the fusion hidden layer. The second part is the single modality energy, which is defined over a CRBM of a single modality m. It consists of unary terms representing the bias of each layer, and a pairwise term which relates the nodes of two layers.

$$\begin{aligned} E_{\mathrm{MC}}(\mathbf{v}_t^{1:M},\mathbf{h}_t^{1:M},\mathbf{h}_t^F|\mathbf{v}_{<t}^{1:M}) = &\underbrace{\sum_m E_{\mathrm{C}}(\mathbf{v}_t^m,\mathbf{h}_t^m|\mathbf{v}_{<t}^m)}_{\text{Unimodal}}\\ &\underbrace{-\sum_j f_{k,t} h_{k,t}^F - \sum_{j,k} h_{j,t}^{1:M} w_{jk}^F h_{k,t}^F}_{\text{Fusion}}. \end{aligned} \tag{12}$$

3.4 Discriminative Restricted Boltzmann Machines (DRBMs)

DRBMs, shown in Fig. 1d, are a natural extension of RBMs which have an additional discriminative term for classification (Larochelle and Bengio 2008). They are based on the model in Larochelle and Bengio (2008). The DRBM defines a probability distribution p_{DR} as a Gibbs distribution (13).

$$\begin{aligned} p_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h}|\mathbf{v}) &= \frac{\exp[-E_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h})]}{Z(\boldsymbol{\theta})}, \\ Z(\boldsymbol{\theta}) &= \textstyle\sum_{\mathbf{y},\mathbf{v},\mathbf{h}} \exp[-E_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h})], \\ \boldsymbol{\theta} &= \begin{bmatrix} \{\mathbf{a},\mathbf{b},\mathbf{s}\} & \text{-bias,} \\ \{W,U\} & \text{-fully connected.} \end{bmatrix} \end{aligned} \tag{13}$$

The probability distribution over the visible layer will follow the same distributions as in (2). The hidden layer $\mathbf{h}$ is defined as a function of the labels y and the visible nodes $\mathbf{v}$. Also, a new probability distribution for the classifier is defined to relate the label y to the hidden nodes $\mathbf{h}$ as in (14).

$$\begin{aligned} p_{\text{DR}}(v_i|\mathbf{h}) &= \mathcal{N}(a_i + \textstyle\sum_j h_j w_{ij}, 1), \\ p_{\text{DR}}(h_j = 1|y_l, \mathbf{v}) &= \sigma(b_j + u_{jl} + \textstyle\sum_i v_i w_{ij}), \\ p_{\text{DR}}(y_l|\mathbf{h}) &= \frac{\exp[s_l + \sum_j u_{jl} h_j]}{\sum_{l^*} \exp[s_{l^*} + \sum_j u_{jl^*} h_j]}. \end{aligned} \tag{14}$$

The new energy function E_{DR} is defined similar to (15),

$$E_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h}) = \underbrace{E_{\text{R}}(\mathbf{v},\mathbf{h})}_{\text{Generative}} \underbrace{- \sum_l s_l y_l - \sum_{j,l} h_j u_{jl} y_l}_{\text{Discriminative}} \tag{15}$$

3.5 Discriminative Conditional Restricted Boltzmann Machines (DCRBMs)

In the same way the RBM can be extended to the DRBM by adding a discriminative term to the model, we can extend the CRBM to the DCRBM (Fig. 1e). DCRBMs are based on the model in Larochelle and Bengio (2008), generalized to account for temporal phenomenon using CRBMs. DCRBMs are a simpler version of the Factored Conditional Restricted Boltzmann Machines (Taylor et al. 2011) and Gated Restricted Boltzmann Machines (Memisevic and Hinton 2007). Both these models incorporate labels in learning representations, however, they use a more complicated potential which involves three way connections into factors. DCRBMs define the probability distribution p_{DC} as a Gibbs distribution (16).

$$\begin{aligned} p_{\text{DC}}(\mathbf{y}_t,\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t};\boldsymbol{\theta}) &= \frac{\exp[-E_{\text{DC}}(\mathbf{y}_t,\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t})]}{Z(\boldsymbol{\theta})}, \\ Z(\boldsymbol{\theta}) &= \textstyle\sum_{\mathbf{y},\mathbf{v},\mathbf{h}} \exp[-E_{\text{DC}}(\mathbf{y}_t,\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t})], \\ \boldsymbol{\theta} &= \begin{bmatrix} \{\mathbf{a},\mathbf{b},\mathbf{s}\} & \text{-bias,} \\ \{A,B\} & \text{-auto regressive,} \\ \{W,U\} & \text{-fully connected.} \end{bmatrix} \end{aligned} \tag{16}$$

The probability distribution over the visible layer will follow the same distributions as in (14). The hidden layer $\mathbf{h}$ is defined as a function of the labels y and the visible nodes $\mathbf{v}$. A new probability distribution for the classifier is defined to relate the label y to the hidden nodes $\mathbf{h}$ (17).

$$\begin{aligned} p_{\text{DC}}(v_{i,t}|\mathbf{h}_t,\mathbf{v}_{<t}) &= \mathcal{N}(c_i + \textstyle\sum_j h_j w_{ij}, 1), \\ p_{\text{DC}}(h_{j,t} = 1|y_t,\mathbf{v}_t,\mathbf{v}_{<t}) &= \sigma(d_j + u_{j,k} + \textstyle\sum_i v_{i,t} w_{ij}), \\ p_{\text{DC}}(y_{l,t}|\mathbf{h}) &= \frac{\exp[s_l + \sum_j u_{j,l} h_j]}{\sum_{l^*} \exp[s_{l^*} + \sum_j u_{j,l^*} h_j]}. \end{aligned} \tag{17}$$

where,

$$\begin{aligned} c_i &= a_{i,t} + \textstyle\sum_p A_{p,i} v_{p,<t}, \\ d_j &= b_{j,t} + \textstyle\sum_p B_{p,j} v_{p,<t}. \end{aligned} \tag{18}$$

The new energy function E_{DC} is defined similar to that of the DRBM (15).

$$\begin{aligned} E_{\text{DC}}(\mathbf{y}_t,\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t}) &= \underbrace{E_{\text{C}}(\mathbf{v}_t,\mathbf{h}_t|\mathbf{v}_{<t})}_{\text{Generative}} \\ &\quad \underbrace{- \sum_{j,l} h_{j,t} u_{jl} y_{l,t} - \sum_l s_l y_{l,t}}_{\text{Discriminative}} \end{aligned} \tag{19}$$

3.6 Multimodal Discriminative Conditional Restricted Boltzmann Machines (MMDCRBMs)

In the same way CRBMs can be extended to MMCRBMs, we can naturally extend DCRBMs to MMDCRBMs. A MMDCRBM combines a collection of unimodal DCRBMs, one for each visible modality. The hidden representations produced by the unimodal DCRBMs are then treated as the visible vector of a single fusion DCRBM. The result is a MMDCRBM model that relates multiple temporal modalities to a classification label. MMDCRBMs define the probability distribution p_{MDC} as a Gibbs distribution (20).

$$p_{\text{MDC}}(\mathbf{y}_t, \mathbf{v}_t^{1:M}, \mathbf{h}_t^{1:M}, \mathbf{h}_t^F | \mathbf{v}_{<t}^{1:M}) = \exp[-E_{\text{MDC}}(\mathbf{y}_t, \mathbf{v}_t^{1:M}, \mathbf{h}_t^{1:M}, \mathbf{h}_t^F | \mathbf{v}_{<t}^{1:M})]/Z(\boldsymbol{\theta}),$$

$$Z(\boldsymbol{\theta}) = \textstyle\sum_{\mathbf{y},\mathbf{v},\mathbf{h}} \exp[-E_{\text{MDC}}(\mathbf{y}_t, \mathbf{v}_t^{1:M}, \mathbf{h}_t^{1:M}, \mathbf{h}_t^F | \mathbf{v}_{<t}^{1:M}),$$

$$\boldsymbol{\theta} = \begin{bmatrix} \{\mathbf{a}^{1:M}, \mathbf{b}^{1:M}, \mathbf{e}, \mathbf{s}\} & \text{-bias,} \\ \{A^{1:M}, B^{1:M}, C^{1:M}\} & \text{-auto regressive,} \\ \{W^{1:M}, W^F, U^{1:M}, U^F\} & \text{-fully connected.} \end{bmatrix} \tag{20}$$

The probability distribution over the visible layer will follow the same distributions as in (14). The hidden layer $\mathbf{h}$ is defined as a function of the labels y and the visible nodes $\mathbf{v}$. A new probability distribution for the classifier is defined to relate the label y to the hidden nodes $\mathbf{h}$ is defined as in (21).

$$\begin{aligned}
&p_{\text{MDC}}(v_{i,t}^m | \mathbf{h}_t^m, \mathbf{v}_{<t}^m) = \mathcal{N}(c_i^m + \textstyle\sum_j h_j^m w_{ij}^m, 1), \\
&p_{\text{MDC}}(h_{j,t}^m = 1 | y_{l,t}, \mathbf{v}_t^m, \mathbf{v}_{<t}^m) = \sigma(d_j^m + u_{jl}^m + \textstyle\sum_i v_{i,t}^m w_{ij}^m), \\
&p_{\text{MDC}}(y_{l,t} | \mathbf{h}_t^m) = \frac{\exp[s_l + \sum_j u_{jl}^m h_{j,t}^m]}{\sum_{l^*} \exp[s_{l^*} + \sum_j u_{jl^*}^m h_{j,t}^m]}, \\
&p_{\text{MDC}}(h_{k,t}^F = 1 | y_{l,t}, \mathbf{h}_t^{1:M}, \mathbf{h}_{<t}^{1:M}) = \sigma(f_k + u_{kl}^F + \textstyle\sum_{m,j} h_{j,t}^m w_{jk}^m), \\
&p_{\text{MDC}}(y_{l,t} | \mathbf{h}_t^F) = \frac{\exp[s_l + \sum_k u_{kl}^F h_{k,t}^F]}{\sum_{l^*} \exp[s_{l^*} + \sum_k u_{kl^*}^F h_{k,t}^F]}.
\end{aligned} \tag{21}$$

where,

$$\begin{aligned}
c_i^m &= a_i^m + \textstyle\sum_p A_{p,i}^m v_{p,<t}^m, \\
d_j^m &= b_j^m + \textstyle\sum_p B_{p,j}^m v_{p,<t}, \\
f_k &= e_k + \textstyle\sum_{m,r} C_{r,k}^m h_{r,<t}^m.
\end{aligned} \tag{22}$$

The new energy function E_{MDC} is defined similar to that of the DRBM (15).

$$E_{\text{MDC}}(\mathbf{y}_t, \mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}) = \underbrace{E_{\text{MC}}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t})}_{\text{Generative}} \underbrace{- \sum_k s_l y_{l,t} - \sum_{k,l} h_{k,t}^F u_{kl}^F y_{l,t} - \sum_{j,l,m} h_{j,t}^m u_{jl}^m y_{l,t}}_{\text{Discriminative}} \tag{23}$$

4 Inference

Classification to perform classification at time t in the MMDCRBM given $\mathbf{v}_{<t}^{1:M}$ and $\mathbf{v}_t^{1:M}$ we use a bottom-up approach,

Algorithm 1: Classification

Input: Multimodal data $\mathbf{v}_t^m$ and history $\mathbf{v}_{<t}^m \quad \forall \quad m \in \{1, 2, \ldots, M\}$
Output: Activity class label per frame y_t

for $l = 1$:*Number of labels L* **do**
  for $m = 1$:*Number of Modalities in M* **do**
    for $j = 1$:*Number of nodes in* $\mathbf{h}^m$ **do**
      Activate the modality's hidden layer: $p_{\text{MDC}}(h_{j,t}^m = 1 | y_{l,t}, \mathbf{v}_t^m, \mathbf{v}_{<t}^m) = \sigma(d_j^m + u_{jl}^m + \sum_i v_{i,t}^m w_{ij}^m)$;
    end
    Compute: $p_{\text{MDC}}(y_{l,t} | \mathbf{h}_t^m) = \frac{\exp[s_l + \sum_j u_{jl}^m h_{j,t}^m]}{\sum_{l^*} \exp[s_{l^*} + \sum_j u_{jl^*}^m h_{j,t}^m]}$;
    Classify the modality label per frame: $\mathbf{y}_t = \arg\max_l p_{\text{MDC}}(y_{l,t} | \mathbf{h}_t^m)$;
  end
  for $k = 1$:*Number of nodes in* $\mathbf{h}^F$ **do**
    Activate the Fusion hidden layer: $p_{\text{MDC}}(h_{k,t}^F = 1 | y_{l,t}, \mathbf{h}_t^{1:M}, \mathbf{h}_{<t}^{1:M}) = \sigma(f_k + u_{kl}^F + \sum_{m,j} h_{j,t}^m w_{jk}^m)$;
  end
  Compute: $p_{\text{MDC}}(y_{l,t} | \mathbf{h}_t^F) = \frac{\exp[s_l + \sum_k u_{kl}^F h_{k,t}^F]}{\sum_{l^*} \exp[s_{l^*} + \sum_k u_{kl^*}^F h_{k,t}^F]}$.
end
Classify the fused representation label per frame: $\mathbf{y}_t = \arg\max_l p_{\text{MDC}}(y_{l,t} | \mathbf{h}_t^F)$;

Algorithm 2: Generation

Input: Activity class label per frame $y_{l,t}$ and initial history $\mathbf{v}_{<N}^m \quad \forall \quad m \in \{1, 2, \ldots, M\}$
Output: Multimodal data $\mathbf{v}_t^m \quad \forall \quad m \in \{1, 2, \ldots, M\}$

for $k = 1$:*Number of nodes in* $\mathbf{h}^F$ **do**
  Activate the Fusion hidden layer: $p_{\text{MDC}}(h_{k,t}^F = 1 | y_{l,t}) = \sigma(f_k + u_{kl}^F)$;
end
for $m = 1$:*Number of Modalities in M* **do**
  for $j = 1$:*Number of nodes in* $\mathbf{h}^m$ **do**
    Activate the modality's hidden layer: $p_{\text{MDC}}(h_{j,t}^m = 1 | y_{l,t}, h_t^F) = \sigma(d_j^m + u_{jl}^m + \sum_k h_t^F w_{jk}^m)$;
  end
  for $i = 1$:*Number of nodes in* $\mathbf{v}^m$ **do**
    $p_{\text{MDC}}(v_{i,t}^m | \mathbf{h}_t^m, \mathbf{v}_{<t}^m) = \mathcal{N}(c_i^m + \sum_j h_j^m w_{ij}^m, 1)$
  end
end

computing a cost for each possible label $\mathbf{y}_t$ then choosing the label with least cost. Ideally we would like the cost for label $\mathbf{y}_t$ to be the free energy $-\log p_{\text{MDC}}(\mathbf{y}_t, \mathbf{v}_t^{1:M} | \mathbf{v}_{<t}^{1:M})$ computed by marginalizing over $\mathbf{h}_{<t}^{1:M}$, $\mathbf{h}_t^{1:M}$, and $\mathbf{h}_t^F$, but this is intractable due to the hidden-hidden edges.

Because this preferred cost is intractable, we use an approximate procedure instead. First, for each modality m, the expected value of $\mathbf{h}_t^m$ is computed according to (21). Then, the cost associated with the candidate label is the free energy in the fusion DCRBM, namely $-\log p_{\text{DC}}(\mathbf{y}_t, \mathbf{h}_t^{1:M} | \mathbf{h}_{<t}^{1:M})$ computed by marginalizing over only $\mathbf{h}_t^F$. Since this marginalization does not involve any hidden-hidden edges, it

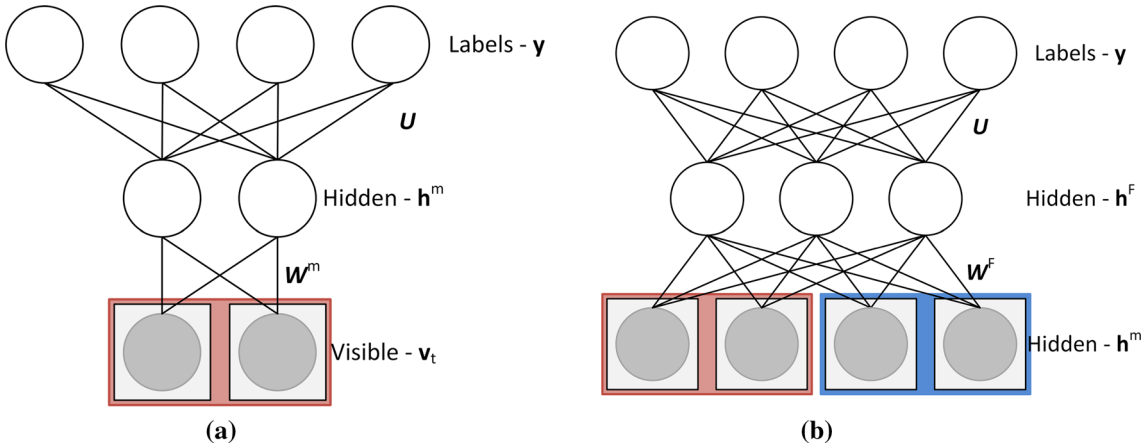

Fig. 2 This figure specifies the inference algorithm. We first classify the unimodal data by activating the corresponding hidden layers $\mathbf{h}_t^m$ as shown in (**a**), followed by classifying the multimodal data by activating the fusion layer $\mathbf{h}_t^F$ as shown in (**b**)

is tractable, because the sum over exponentially many terms can be algebraically eliminated. Figure 2 illustrates our inference. The details of the classification algorithm is shown in Algorithm 1.

Generation to perform unimodal generation for modality m at time t given $\mathbf{v}_{<t}^m$ and $\mathbf{y}_t$ we initialize $\mathbf{v}_t^m$ to $\mathbf{v}_{t-1}^m$ then sample the distribution p_{DC} (17) using Gibbs sampling. Each Gibbs cycle samples $p_{\text{DC}}(\mathbf{h}_t^m|\mathbf{y}_t, \mathbf{v}_t^m, \mathbf{v}_{<t}^m)$ then sampling $p_{\text{DC}}(\mathbf{v}_t^m|\mathbf{h}_t^m, \mathbf{v}_{<t}^m)$. In the last Gibbs cycle, $\mathbf{h}_t^m$ is assigned its expected value according to the distribution, instead of being sampled, we used 50 Gibbs cycles. The details of the generation algorithm is shown in Algorithm 2.

5 Learning

Learning our model is done using contrastive divergence (CD) (Hinton 2002), where $\langle\cdot\rangle_{data}$ is the expectation with respect to the data distribution and $\langle\cdot\rangle_{recon}$ is the expectation with respect to the reconstructed data. The learning is done using two steps a bottom-up pass and a top-down pass using sampling equations from (21).

Bottom-up the reconstruction is generated by first sampling the unimodal layers $p(h_{t,j}^m = 1|\mathbf{v}_t^m, \mathbf{v}_{<t}^m, y_l)$ for all the hidden nodes in parallel. This is followed by sampling the fusion layer $p(h_{t,k}^F = 1|\mathbf{h}_t^{1:M}, \mathbf{h}_{<t}^{1:M}, y_l)$. This is done using the classification algorithm Algorithm 1.

Top-down The unimodal layer is generated using the activated fusion layer $p(h_{t,j}^m = 1|\mathbf{h}_t^F, y_l)$. This is followed by sampling the visible nodes $p(v_{t,i}^m|\mathbf{h}_t^m, \mathbf{v}_{<t}^m)$ for all the visible nodes in parallel. The gradient updates are described in (24). This is done using the generation algorithm Algorithm 2.

$$
\begin{aligned}
\Delta a_i &\propto \langle v_i^m\rangle_{data} &&- \langle v_i^m\rangle_{recon},\\
\Delta b_j &\propto \langle h_j^m\rangle_{data} &&- \langle h_j^m\rangle_{recon},\\
\Delta e_k &\propto \langle h_k^F\rangle_{data} &&- \langle h_k^F\rangle_{recon},\\
\Delta s_l &\propto \langle y_l\rangle_{data} &&- \langle y_l\rangle_{recon},\\
\Delta A_{p,i,<t}^m &\propto v_{k,<t}^m(\langle v_{i,t}^m\rangle_{data} &&- \langle v_{i,t}^m\rangle_{recon}),\\
\Delta B_{p,j,<t}^m &\propto v_{i,<t}^m(\langle h_{j,t}^m\rangle_{data} &&- \langle h_{j,t}^m\rangle_{recon}),\\
\Delta C_{r,k,<t}^m &\propto h_{j,<t}^m(\langle h_{k,t}^F\rangle_{data} &&- \langle h_{j,t}^F\rangle_{recon}),\\
\Delta w_{i,j}^m &\propto \langle v_i^m h_j^m\rangle_{data} &&- \langle v_i^m h_j^m\rangle_{recon},\\
\Delta w_{j,k}^F &\propto \langle h_j^m h_k^F\rangle_{data} &&- \langle h_j^m h_k^F\rangle_{recon},\\
\Delta u_{j,l}^m &\propto \langle y_l h_j^m\rangle_{data} &&- \langle y_l h_j^m\rangle_{recon},\\
\Delta u_{k,l}^F &\propto \langle y_k h_l^F\rangle_{data} &&- \langle y_k h_l^F\rangle_{recon}.
\end{aligned}
\tag{24}
$$

6 Experiments

In Sect. 6.1 we describe the datasets we use for evaluation; In Sect. 6.2 we specify the implementation details and model parameters selection; In Sect. 6.3 we explain how we selected the model parameters; Finally, in Sect. 6.4 we present our results.

6.1 Datasets

We focus our analysis on temporal multimodal datasets from raw sensor data. In the literature we found some relevant datasets, we decided to evaluate our approach on two realistic datasets and three toy datasets that would highlight the contribution of our approach.

The two realistic datasets are: *The Tower Game* dataset (Salter et al. 2015) where its an interaction between two humans with goal of classifying entrainment. The dataset is captured using a Kinect sensor. For this dataset we evaluate the model classification and generating accuracy using mocap-mocap multimodal data; *ChaLearn* dataset (Escalera et al. 2014). This dataset is captured in a similar manner to the

Tower Game Dataset except that they provide audio. For this dataset we evaluate the model classification and generating accuracy using mocap-audio multimodal data.

The three toy datasets are: *AVEC* (Schuller et al. 2011) is an audio-visual dataset for single person affect analysis. *AVLetters* (Matthews et al. 2002), consists of 10 speakers uttering the letters A to Z, three times each. *CUAVE* (Patterson et al. 2002), consists of 36 speakers uttering the digits 0 to 9. AVEC, AVLetters, and CUAVE, are relatively simple datasets for the task we address.

Other relevant datasets include the Multimodal Dyadic Behavior dataset (Rehg et al. 2013), which focuses on analyzing dyadic social interactions between adults and children in a developmental context. The dataset was not fully released. *Mimicry database* (Sun et al. 2011) which focuses on studying social interactions between humans with the aim of analyzing mimicry in human-human interactions. This dataset was collected in an unstructured format where the two humans talk to each other about different subjects. We were unable to gain access to the dataset due to being a non-educational institute.

6.1.1 Realistic Datasets

The Tower Game dataset (Salter et al. 2015) is a simple game of tower building often used in social psychology to elicit different kinds of interactive behaviors from the participants. It is typically played between two people working with a small fixed number of simple toy blocks that can be stacked to form various kinds of towers. We choose these tower games as they force the players to engage and communicate with each other in order to achieve the objectives of the game, thereby evoking behaviors such as *joint-attention* and *entrainment* from the participants. The game, due to its simplicity, allows for total control over the variables of an interaction. Due to the small number of blocks involved, the number of potential moves (actions) is limited. Also since the game involves interacting with physical objects, *joint-attention* is mediated through concrete objects. Furthermore, only two players are involved, ensuring that we can stay in the realm of dyadic interactions. The data consists of 112 videos which were divided into 1213, 10-s, segments indicating the presence or absence of these behaviors in each segment. Entrainment is the alignment in the behavior of two individuals and it involves simultaneous movement, tempo similarity, and coordination. Each measure was rated using a low, medium, high measure for the entire 10 s segment. 70% of that data was used for training and 30% were used for testing. In this dataset we call each person's skeletal data a modality, where our goal is to model mocap-mocap representations. **ChaLearn dataset** (Escalera et al. 2014) consists of a set of italian gestures, featured by a challenge in 2014. The dataset was designed to evaluate user independent continuous Gesture Recognition performance. The dataset consists of 13,858 gestures from a vocabulary of 20 Italian cultural signs performed by 27 unique users. The list of Italian Gestures in the dataset: vattene, ok, vieniqui, cosatifarei, perfetto, basta, furbo, prendere, cheduepalle, noncenepiu, chevuoi, fame, daccordo, tantotempo, seipazzo, buonissimo, combinato, messidaccordo, freganiente, sonostufo. The dataset was recorded by Kinect sensors, and it includes skeleton model, user mask, RGB, and depth images. The dataset consists of 450 development, 250 validation, and 240 test videos. Each gesture is labeled using ground truth gesture type and its start and end timestamps. There are a total of 7754 instances for development, 3362 for validation, and 2742 for testing. The dataset was featured by ChaLearn 2014 Looking at People competition's Track 3: Gesture Recognition. The emphasis of the gesture recognition track was on multi-modal automatic learning of a set of 20 gestures performed by several different users, with the aim of performing user independent continuous gesture localization. We followed the setup provided by Neverova et al. (2014) by using their augmented dataset which contains audio. In this dataset our goal is to model mocap-audio representations.

6.1.2 Toy Datasets

To compare against the prior work of Amer et al. (2014); Ngiam et al. (2011) we evaluate our approach on three toy datasets, AVEC (Schuller et al. 2011), AVLetters (Matthews et al. 2002), and CUAVE datasets (Patterson et al. 2002) which were used in their experiments. **AVEC dataset** (Schuller et al. 2011) is an audio-visual dataset for single person affect analysis. The dataset involves users interacting with emotionally stereotyped virtual characters operated by a human. The visual data contains mainly the face of the user interacting with the character. The Audio data consists of recordings of utterances of the user and is synchronized with the video. The dataset has been annotated with binary labels for four different affective dimensions - Activation, Expectation, Power and Valence. We use the AVEC dataset to compare against (Ramirez et al. 2011; Glodek et al. 2011). The dataset is divided into two sets, 31 sequences for training and 32 sequences for testing. **AVLetters dataset** (Matthews et al. 2002), consists of 10 speakers uttering the letters A to Z, three times each. The dataset also provides pre-extracted 60×80 patches of lip regions along with audio features (MFCC features of 483 dimensions). The dataset is divided into two sets, 2/3 of the sequences for training and 1/3 for testing. **CUAVE dataset** (Patterson et al. 2002), consists of 36 speakers uttering the digits 0 to 9. The dataset provides the aligned face of each speaker of size 75×50, as well as the audio spectrogram and MFCC features of dimensionality 534. The dataset is divided into two sets, 1/2 for training and

1/2 for testing. We follow the same experimental setup as in Ngiam et al. (2011).

6.2 Implementation Details

For ChaLearn and Tower Game datasets pre-processing the mocap data, we followed the same approach as Neverova et al. (2014) by forming a body centric transformation of the skeletons generated by the Kinect sensors. For mocap data we use the 11 joints from the upper body of the two players since the tower game almost entirely involves only upper body actions as well as gestures are done using the upper body. We used the raw joint locations normalized with respect to a selected origin point. We use the same descriptor provided by Neverova et al. (2014), Zanfir et al. (2013). The descriptor consists of 84 dimensions based on the normalized joints location, inclination angles formed by all triples of anatomically connected joints, azimuth angles between projections of the second bone and the vector on the plane perpendicular to the orientation of the first bone, bending angles between a basis vector, perpendicular to the torso, and joint positions.

For the audio component of ChaLearn, pre-processing the audio stream, we followed the same approach as Neverova et al. (2014) by using feature learning within a convolution architecture for audio. First, they apply a short-time Fourier transform on the raw audio signal to obtain an audio spectrogram. Second, they transform the spectrogram to the Mel-scale to produce 40 filterbanks. Finally, they input the filterbanks to a one-layer convolutional network in combination with two fully-connected layers resulting in a 40 dimensional feature. For the ChaLearn dataset we trained one multi-class model on the 20 gestures and background. For the Tower Game dataset we trained three multi-class models, one for each of the labels, Tempo Similarity, Coordination, and Simultaneous Movement since they could co-occur with three different values {low, medium, high}.

AVEC dataset comes with pre-computed audio and video features; refer to Schuller et al. (2011) for details. We apply PCA on the extracted features and reduce each of the audio to 100 dimensions and video features to 32 dimensions. For each modality we choose a CRBM with a temporal order $N = 5$, with the first hidden layer being over-complete consisting of 150 nodes, and the multimodal fusion layer consisting of 300 nodes. As with the AVLetters dataset, following the same setup as in Ngiam et al. (2011), we reduce the dimensionality of the audio features to 100 dimensions using PCA whitening and the video features (lip region) to 32 dimensions. Similarly for CUAVE dataset, we reduce the dimensionality of the audio features to 100 dimensions using PCA whitening and the video features (lip region) to 32 dimensions. For AVEC we trained 4 different classifiers with labels Activity, Expectation, Power, and Valence. For

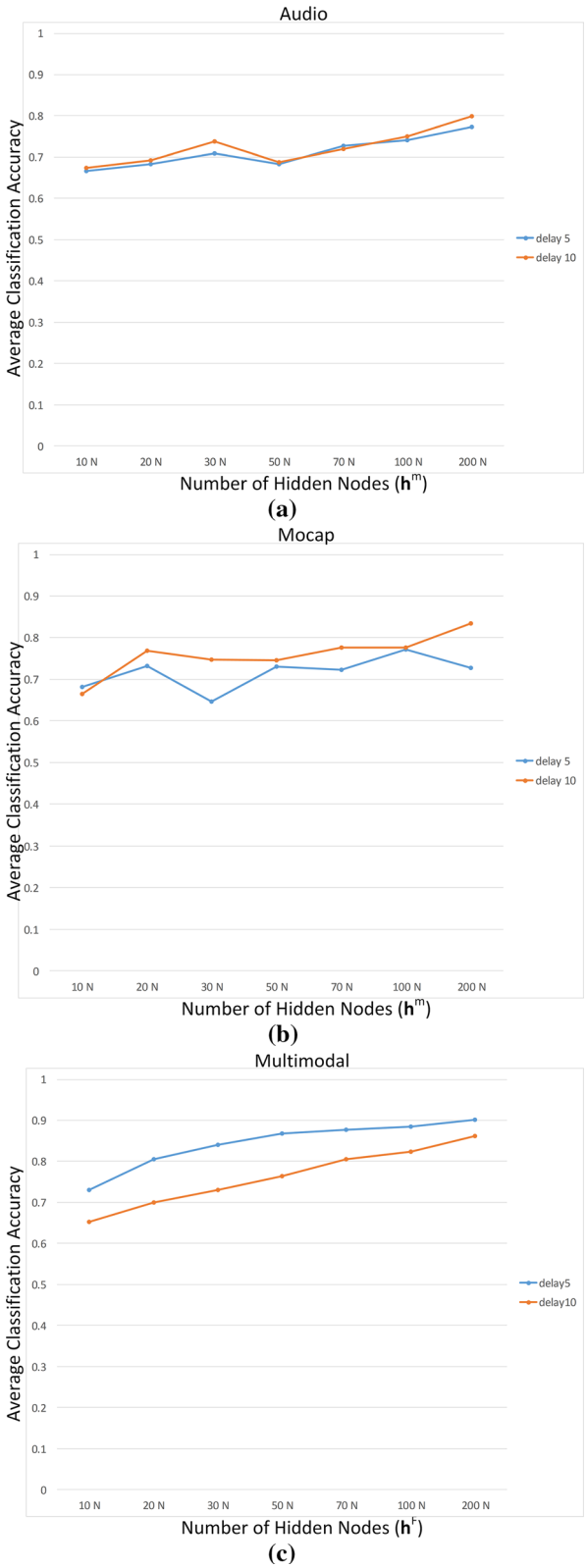

Fig. 3 This figure shows the sensitivity of our model's average classification accuracy to the number of hidden nodes and auto-regressive edges on the ChaLearn dataset. **a**–**c** show the sensitivity of our audio, mocap, multimodal classifiers respectively to the number of hidden nodes and auto-regressive edges

AVLetters we trained multi-class classifier with 26 classes and for CUAVE a multi-class classifier with 10 classes.

6.3 Model Selection

We tuned our model parameters on ChaLearn dataset. For selecting the model parameters we used a grid search. We varied the number of hidden nodes per layer in the range of {10, 20, 30, 50, 70, 100, 200}, as well as the auto-regressive nodes in the range of {5, 10}, resulting a total of 2744 trained models using the development set and used them to classify the validation set. Figure 3 shows the average classification accuracy of the different models (per hidden layer) and the different delays. The best performing model on ChaLearn has the following configuration:

$$\begin{array}{lllll} \text{Mocap:} & v & = 84, & h^m = 30, & < t = 10, \\ \text{Audio:} & v & = 40, & h^m = 200, & < t = 5, \\ \text{Multimodal:} & h^{1:M} & = 230, & h^F = 200, & < t = 5. \end{array}$$

The best performing model on Tower Game Dataset has the following configuration:

$$\begin{array}{lllll} \text{Mocap-1:} & v & = 84, & h^m = 30, & < t = 10, \\ \text{Mocap-2:} & v & = 84, & h^m = 30, & < t = 10, \\ \text{Multimodal:} & h^{1:M} & = 60, & h^F = 200, & < t = 5. \end{array}$$

The best performing model on AVEC, AVLetters and CUAVE Dataset has the following configuration:

$$\begin{array}{lllll} \text{Visual:} & v & = 32, & h^m = 150, & < t = 10, \\ \text{Audio:} & v & = 100, & h^m = 150, & < t = 5, \\ \text{Multimodal:} & h^{1:M} & = 300, & h^F = 300, & < t = 5. \end{array}$$

6.4 Quantitative Results

To evaluate our model we use three different metrics: (1) classification, (2) generation, and (3) localization. Note that for the ChaLearn dataset only the localization results were reported by other performers (Escalera et al. 2014). We are the first to report the generation results on this dataset since all the previous work done was using discriminative classifiers. As for the Tower Game dataset (Salter et al. 2015) we report classification accuracy as well as generation error. For AVEC (Schuller et al. 2011), AVLetters (Matthews et al. 2002), and CUAVE (Patterson et al. 2002) we report average classification accuracy which was the commonly used metric of performance.

6.4.1 Classification

In ChaLearn Dataset we evaluated our average classification accuracy with respect to different size of training sets. Figure 4 shows our sensitivity with respect of the amount

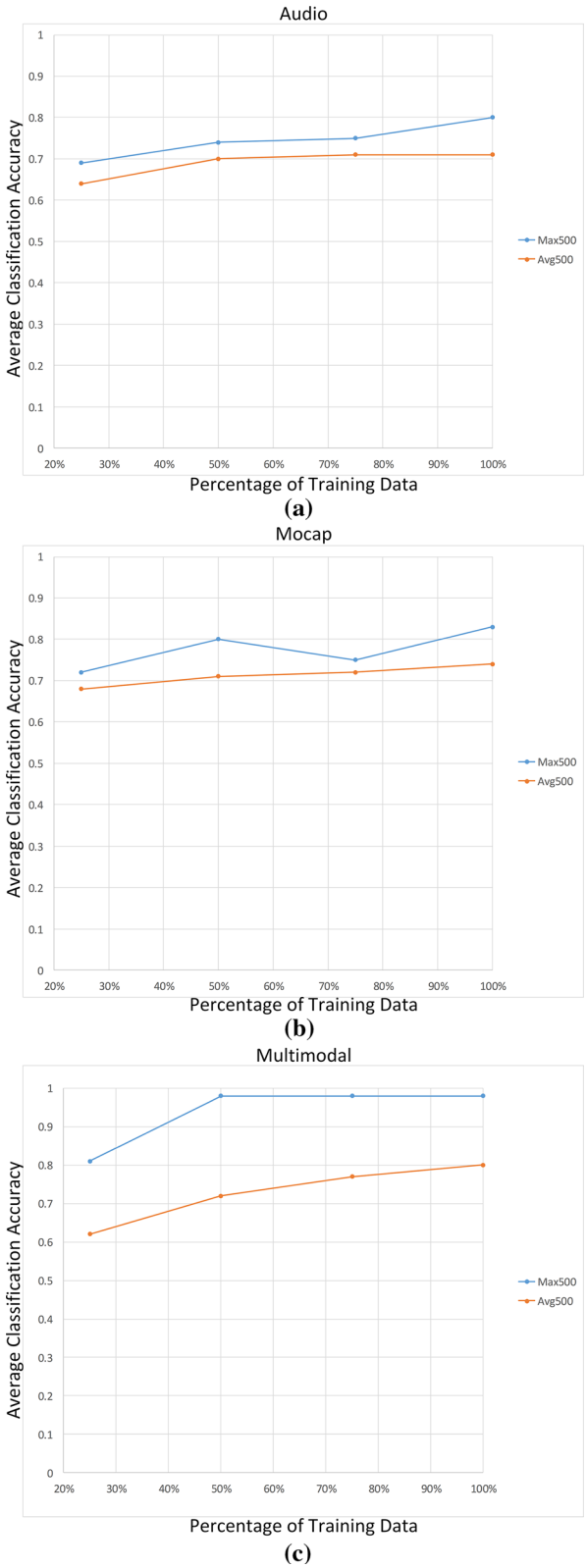

Fig. 4 This figure shows the sensitivity of our model's average classification accuracy to amount of training data on the ChaLearn dataset. **a–c** show the sensitivity of our audio, mocap, multimodal classifiers respectively to more training data

Table 1 Average classification accuracy on simultaneous movement

Classifier/Annot.	A1	A2	A3	A4	A5	A6	All
Random Guess	33.3	33.3	33.3	33.3	33.3	33.3	33.3
SVM-Raw Skel.257400D	45.2	37.5	32.6	38.4	37.1	50.5	39.5
SVM-PCA 100D	42.8	21.3	37.4	39.4	31.8	62.7	47.8
SVM-BoW 100D	33.2	36.6	35.5	36.9	39.3	50.0	44.3
SVM-BoW 300D	38.5	33.5	40.8	41.6	41.2	50.1	42.8
SVM-RBM 84D	43.5	31.0	42.2	38.7	41.9	56.4	43.2
CRF-RBM 84D	43.9	32.2	43.9	40.0	43.1	57.2	43.8
MMDRBM 84D	49.1	34.8	44.0	40.2	45.5	66.5	44.0
SVM-CRBM (Amer et al. 2014) 84D	48.5	33.5	43.5	39.1	44.4	60.0	43.5
CRF-CRBM (Amer et al. 2014) 84D	50.0	33.7	44.3	43.2	45.1	62.1	45.7
MMDCRBM 84D-Ours	55.1	40.8	48.6	45.4	50.0	70.5	49.2

Table 2 Average classification accuracy on coordination

Classifier/Annot.	A1	A2	A3	A4	A5	A6	All
Random Guess	33.3	33.3	33.3	33.3	33.3	33.3	33.3
SVM-Raw Skel.257400D	59.0	54.6	46.9	73.3	79.2	50.6	52.2
SVM-PCA 100D	58.8	48.0	52.3	79.7	83.3	44.6	58.2
SVM-BoW 100D	52.2	63.6	40.1	58.0	73.3	47.4	55.8
SVM-BoW 300D	60.3	64.9	35.1	59.5	74.9	51.3	47.5
SVM-RBM 84D	61.6	50.0	41.5	68.9	82.2	50.4	58.9
CRF-RBM 84D	64.2	54.8	47.2	73.0	85.6	52.4	59.2
MMDRBM 84D	71.6	60.2	48.9	79.4	87.6	53.0	59.8
SVM+CRBM (Amer et al. 2014) 84D	70.5	47.3	43.0	78.2	87.9	53.8	59.4
CRF+CRBM (Amer et al. 2014) 84D	72.4	60.1	48.7	79.1	89.1	55.8	61.6
MMDCRBM 84D-Ours	85.7	68.2	57.2	82.5	89.3	55.9	62.0

Table 3 Average classification accuracy on tempo similarity

Classifier/Annot.	A1	A2	A3	A4	A5	A6	All
Random Guess	33.3	33.3	33.3	33.3	33.3	33.3	33.3
SVM-Raw 257400D	68.7	44.6	53.5	82.4	72.0	71.7	59.3
SVM-PCA 100D	64.9	50.5	51.8	86.2	82.2	81.3	72.8
SVM-BoW 100D	65.1	45.6	40.5	58.0	73.3	47.4	65.6
SVM-BoW 300D	57.0	41.3	46.1	77.2	58.1	69.6	54.4
SVM-RBM 84D	66.4	50.8	44.2	82.7	75.3	73.0	68.0
CRF-RBM 84D	68.1	51.7	49.3	86.2	76.2	79.9	70.6
MMDRBM 84D	69.6	51.2	51.8	87.0	77.9	80.2	71.3
SVM-CRBM 84D	68.0	51.0	45.6	86.1	77.0	77.4	71.7
CRF-CRBM 84D	69.9	52.9	52.1	87.4	77.1	81.2	76.3
MMDCRBM 84D-Ours	71.9	55.8	54.0	88.8	85.5	83.0	76.5

of training data used, we report the average over all classifiers 2744 models trained as well as the best performing classifier. Our approach was able to achieve relatively good results using only 25% of the training data. We achieve the best average classification accuracy results on the multimodal layer using only 50% of the data, which shows how powerful our model is learning from the first set of data. Our best configuration achieves 80.5%, 83.1%, and 98.5% average classification accuracy for audio, mocap, and multimodal respectively.

In the Tower game dataset label can be assigned *low*, *medium* or *high*. The data is split into two sets, a training set consisting of 70% of the instances, and a test set consisting of the remaining 30%. We performed a 5 fold cross

Table 4 Classification accuracy on CUAVE dataset

Model	Accuracy
Discrete Cosine Transform (Gurban and Thiran 2009)	64
Fused Holistic+Patch (Lucey and Sridharan 2006)	77.08
Visemic AAM (Papandreou et al. 2009)	83
SVM-RBM (Ngiam et al. 2011)	66.7
CRF-RBM (Amer et al. 2014)	68.6
CRF-CRBM (Amer et al. 2014)	69.1
MMDRBM	71.8
MMDCRBM	74.6

Table 5 Classification accuracy on AVLetters dataset

Model	Accuracy
Multiscale Spatial Analysis (Matthews et al. 2002)	44.6
LBP (Zhao and Barnard 2009)	58.85
SVM-RBM (Ngiam et al. 2011)	59.2
CRF-RBM (Amer et al. 2014)	63.8
CRF-CRBM (Amer et al. 2014)	67.1
MMDRBM	69.0
MMDCRBM	72.5

Table 6 Average classification accuracy on AVEC dataset (Schuller et al. 2011)

Model	mean Accuracy
Baseline-RAW (Schuller et al. 2011)	65.27
SVM-RAW (Late Fusion) (Ramirez et al. 2011)	70.55
LDCRF-RAW (Late Fusion) (Ramirez et al. 2011)	75.40
PLS-SVM (Late Fusion) (Siddiquie et al. 2013)	67.37
CRF-RAW (Late Fusion) (Siddiquie et al. 2013)	69.97
HCRF-RAW (Late Fusion) (Siddiquie et al. 2013)	69.90
JHCRF-RAW (Siddiquie et al. 2013)	71.85
CRF-RBM (Amer et al. 2014)	68.4
CRF-CRBM (Amer et al. 2014)	70.8
MMDRBM	74.2
MMDCRBM	76.1

validation to guarantee unbiased results. Tables 1, 2, and 3 shows our average classification accuracy on the Tower Game Dataset using different features and baselines combinations as well as the results from our MMDCRBM model. The evaluation is done with respect to the six annotators $\{A_1, A_2, \ldots, A_6\}$ as well as the mean annotation. We compare our approach against the baselines presented in Salter et al. (2015), where they extracted a set of first order static and dynamic handcrafted skeleton features. The static features are computed per frame. The features consist of relationships between all pairs of joints of a single actor, and the relationships between all pairs of joints of both the actors. The dynamic features are extracted per window (a set of 300 frames). In each window, they compute first and second order dynamics (velocities and accelerations) of each joint, as well as relative velocities and accelerations of pairs of joints per actor, and across actors. The dimensionality of their static and dynamic features is (257400 D). To reduce their dimensionality they used Principle Component Analysis (PCA) (100 D), Bag-of-Words (BoW) (100 and 300 D) (Niebles et al. 2008). We can see that the MMDCRBM model outperforms all the other models for each of the three measures across all annotators, thereby demonstrating its effectiveness on detecting these entrainment measures. Furthermore, the MMDCRBM model outperforms the PCA and BoW based features which are derived from the high dimensional handcrafted features, demonstrating its ability to learn a rich representation starting from the raw skeleton features.

Table 7 Localization accuracy on ChaLearn dataset

Rank-approach	Jaccard index (J)
#1—Neverova et al. (2014)	0.870
#2—Neverova et al. (2014)	0.850
#3—Monnier et al. (2014)	0.834
#4—Chang (2014)	0.827
#5—Peng et al. (2014)	0.792
#6—Pigou et al. (2014)	0.789
#7- Ours MMDCRBM (audio/mocap)	**0.788**
#8—Wu (2014)	0.787
#9—MMDRBM (audio/mocap)	0.752
#10—Camgoz et al. (2014)	0.747
#11—Evangelidis et al. (2014)	0.745
#12—CRF-CRBM (audio/mocap) Amer et al. (2014)	0.701
#13—Undisclosed	0.689
#14—Ours DCRBM(mocap only)	**0.65**
#15—Chen et al. (2014)	0.649
#16—SVM-CRBM(audio/mocap) Amer et al. (2014)	0.646
#17—CRF-CRBM(mocap Only) Amer et al. (2014)	0.641
#18—SVM-CRBM(mocap Only) Amer et al. (2014)	0.638
#19—DRBM(mocap Only)	0.632

For CUAVE dataset (Patterson et al. 2002) Table 4 shows the classification performance for visual speech recognition. Note that the models (Gurban and Thiran 2009; Lucey and Sridharan 2006; Papandreou et al. 2009), use a pre-processing step that is substantially more complex than ours. In our case, we use the same pre-processing as in Ngiam et al. (2011) which extracts bounding boxes ignoring orientation and perspective changes. Table 5 shows the classification performance for visual speech recognition on the AVLetters dataset (Cox et al. 2008). Our hybrid model shows a substantial improvement over the state-of-the-art which include the hand-engineered features (Matthews et al. 2002; Zhao and Barnard 2009) as well as the staged hybrid models CRF-CRBM (Amer et al. 2014) model and SVM-RBM (Ngiam et al. 2011). In AVEC dataset we evaluated the average classification accuracy in Table 6. Again, we can see that our MMDCRBM model outperformed all other models, followed by the staged CRF-CRBM and CRF-RBM models (Amer et al. 2014).

6.4.2 Localization

We used the Jaccard Index to evaluate the localization performance on ChaLearn dataset for the continuous sequences provided in the test set. The Jaccard Index was proposed by the challenge organizers to standardize a way for comparison between approaches and is defined in (25). $A_{s,n}$ is the ground truth label for gesture n in sequence s, and $B_{s,n}$ is the predicted label. This index provides an evaluation of the area of overlap as well as predicting the correct label. Table 7 shows the localization results on the ChaLearn dataset. Note that for localization we used a simple scanning window approach with no smoothing or post-processing and we were able to achieve 7th position using (audio and mocap) and 14th position using mocap only. Note that our iteratively trained model out performs the staged models CRF-CRBM (Amer et al. 2014) which further confirms that iterative training improves the representation learned. Also note that our approach uses only 10% of the parameters used by Neverova et al. (2014).

$$J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}} \tag{25}$$

6.4.3 Generation

We evaluated our DCRBM mocap generation error. Given the class label and initial history data and our goal is to generate the full visible layer (i.e. the raw features) for that label. This task allows us to visualize what the classifier has learned. We sample the hidden representation h_t^m and then generate frames using 50 Gibbs cycles where the last Gibbs cycle samples the mean of the hidden representation h_t^m instead of a sample. representation. The generation error is calculated using 26 and is averaged over 100 instances.

For the Tower Game dataset the sequence size is 300 frames so we vary the generated data window size from 0

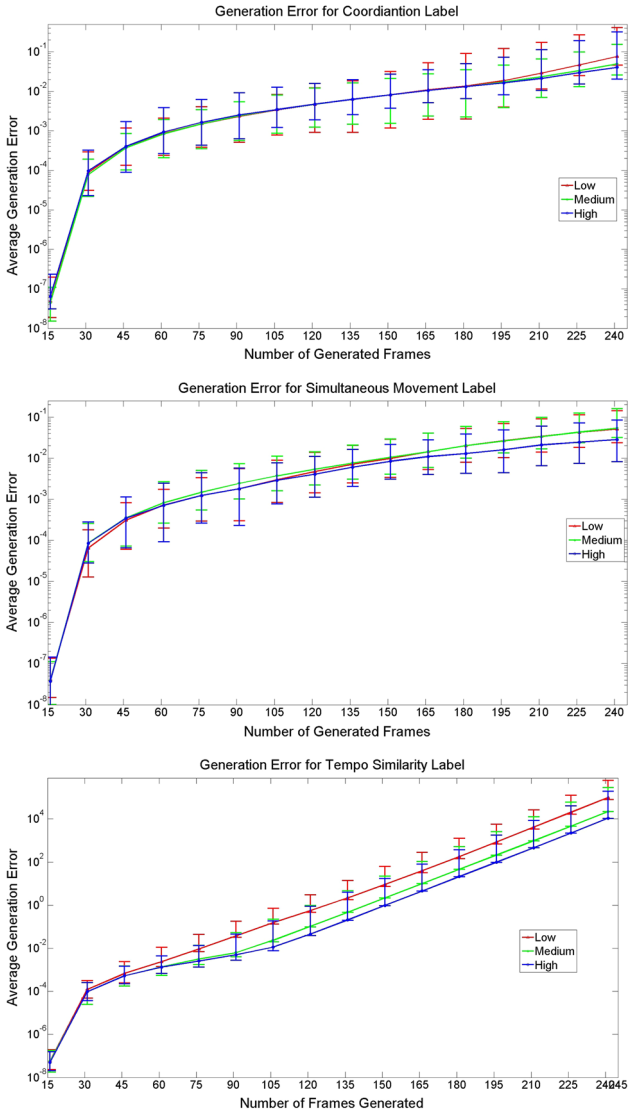

Fig. 5 Unimodal average generation error on the tower game dataset. The results are averaged over 100 instances per class. The average generation error (*y-axis*) for the full visible layer by varying the generated window size (*x-axis*)

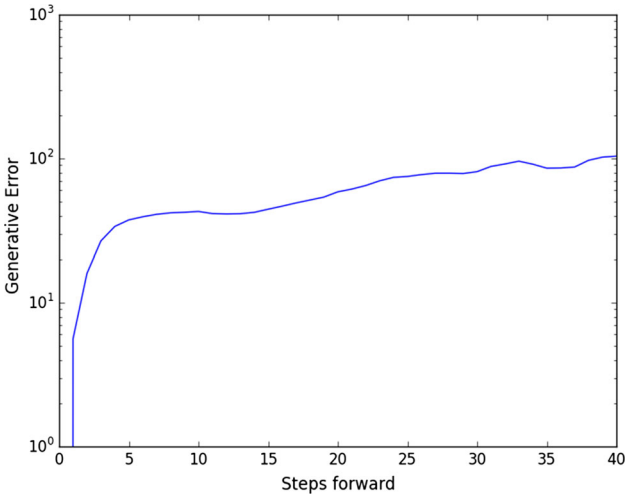

Fig. 6 Unimodal average generation error on ChaLearn dataset. The results are averaged over all 21 classes and over 100 instances per class. The average generation error (*y-axis*) for the full visible layer by varying the generated window size (*x-axis*)

to 300 frames as shown in Fig. 5. We can see that the generation error is relatively low (<0.1) in all cases (except for Tempo Similarity. Tempo Similarity measures the similarity in the rate of the motion of the two players, and when data from both the players is missing generating their raw features based on whether their rate of motion is similar is extremely under constrained, when generating the entire visible layer data) demonstrating the effectiveness of DCRBM model for generating data. Also, the error is similar across different levels (strengths) for each measure indicating that the model is relatively stable. For ChaLearn dataset the average validation sequence size is 40 frames, so we vary the generated window size from 0 to 40 as show in Fig. 6. The error is a bit higher for the ChaLearn dataset since the gestures are structured. We visualized the generated frames and found that the model identifies the most discriminative pose of the gesture and locks onto it. Finally, the error increases with the length of the generated sequence, which is expected as the possibility of variation in the ground-truth sequences increases with length. Note that our approach is the only generative approach that was evaluated on either of the datasets.

$$\text{Generation Error} = \left(\frac{\|\mathbf{v}_{\text{Generated}} - \mathbf{v}_{\text{Groundtruth}}\|}{\|\mathbf{v}_{\text{Groundtruth}}\|} \right)^2 \tag{26}$$

7 Conclusion

We have proposed a hybrid model comprising of temporal generative and discriminative models for classifying sequential data from multiple heterogeneous modalities. Our research resulted in the development of two main models: the DCRBM, which combines temporal, discriminative, and generative concerns into a single RBM-based model; and the MMDCRBM, which fuses multiple DCRBMs, enabling the learning of a rich fused feature representation combining multiple modalities. We employ a energy based temporal generative model which enables us to learn a joint representation to model the short-term temporal characteristics, while also allowing us to handle missing data. An extensive experimental evaluation on two different realistic datasets and three toy datasets demonstrates the superiority of our approach over the state-of-the-art. These models are competitive with feedforward neural networks while using much fewer parameters and being generative. Furthermore we man-

aged to reduce the number of parameters to 10% of the best performing method using only two modalities.

Acknowledgements We would like to thank Dr. Natalia Nevrova for providing the features preprocessing code for the ChaLearn dataset, and Dr. Graham Taylor for his insightful feedback and discussions. This work is supported by DARPA W911NF-12-C-0001 and the Air Force Research Laboratory (AFRL). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Amer, M., Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2014). Multimodal fusion using dynamic hybrid models. In *WACV*.

Bengio, Y. (2009). Learning deep architectures for ai. In *FTML*.

Camgoz, N., Kindiroglu, A., & Akarun, L. (2014). Gesture recognition using templatebased random forest classifiers. In *ECCV*.

Chang, J. (2014). Nonparametric gesture labeling from multi-modal data. In *ECCV-W*.

Chen, G., Clarke, D., Giuliani, M., Weikersdorfer, D., & Knoll, A. (2014). Multi-modality gesture detection and recognition with unsupervision, randomization and discrimination. In *ECCV-W*.

Cox, S., Harvey, R., Lan, Y., & Newman, J. (2008). The challenge of multispeaker lip-reading. In *AVSP*.

Druck, G., & McCallum, A. (2010). High-performance semi-supervised learning using discriminatively constrained generative models. In *ICML*.

Escalera, S., Baro, X., Gonzalez, J., Bautista, M., Madadi, M., Reyes, M., Ponce, V., Escalante, H., Shotton, J., & Guyon, I. (2014). Chalearn looking at people challenge 2014: Dataset and results. In *ECCV-W*.

Evangelidis, G., Singh, G., & Horaud, R. (2014). Continuous gesture recognition from articulated poses. In *ECCV-W*.

Fujino, A., Ueda, N., & Saito, K. (2008). Semi-supervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle. In *TPAMI*.

Garg, N., & Henderson, J. (2011). Temporal restricted Boltzmann machines for dependency parsing. In *ACL*.

Glodek, M., et al. (2011). Multiple classifier systems for the classification of audio-visual emotional states. In *ACII*.

Gurban, M., & Thiran, J. P. (2009). Information theoretic feature extraction for audio-visual speech recognition. *IEEE Transactions on Signal Processing*, *57*, 4765–4776.

Hausler, C., & Susemihl, A. (2012). Temporal autoencoding restricted Boltzmann machine. In *CoRR*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. In *NC*.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. In *NC*.

Larochelle, H., & Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. In *ICML*.

Lewandowski, N. B., Bengio, Y., & Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *ICML*.

Li, X., Lee, T., & Liu, Y. (2011). Hybrid generative-discriminative classification using posterior divergence. In *CVPR*.

Lucey, P., & Sridharan, S. (2006). Patch based representation of visual speech. In *HCSnet workshop on the use of vision in human-computer interaction*.

Matthews, I., et al. (2002). Extraction of visual features for lipreading. In: *TPAMI*.

Memisevic, R. & Hinton, G. E. (2007). Unsupervised learning of image transformations. In *CVPR*.

Mohamed, A. R., & Hinton, G. E. (2009). Phone recognition using restricted Boltzmann machines. In *ICASSP*.

Monnier, C., German, S., & Ost, A. (2014). A multi-scale boosted detector for efficient and robust gesture recognition. In *ECCV-W*.

Neverova, N., Wolf, C., Taylor, G., & Nebout, F. (2014). Moddrop: Adaptive multi-modal gesture recognition. In *PAMI*.

Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2014). Multi-scale deep learning for gesture detection and localization. In *ECCV-W*.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. (2011). Multimodal deep learning. In *ICML*.

Niebles, J., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, *79*(3), 299–318.

Papandreou, G., Katsamanis, A., Pitsikalis, V., & Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. In *TASLP*.

Patterson, E., et al. (2002). Cuave: A new audio-visual database for multimodal human-computer interface research. In *ICASSP*.

Peng, X., Wang, L., & Cai, Z. (2014). Action and gesture temporal spotting with super vector representation. In *ECCV-W*.

Perina, A., et al. (2012). Free energy score spaces: Using generative information in discriminative classifiers. In *TPAMI*.

Pigou, L., Dieleman, S., & Kindermans, P. J. (2014). Sign language recognition using convolutional neural networks. In *ECCV-W*.

Ramirez, G., Baltrusaitis, T., & Morency, L. P. (2011). Modeling latent discriminative dynamic of multi-dimensional affective signals. In *ACII*.

Ranzato, M. A., et al. (2011). On deep generative models with applications to recognition. In *CVPR*.

Rehg, J. M., et al. (2013). Decoding children's social behavior. In *CVPR*.

Salakhutdinov, R., & Hinton, G. E. (2006). Reducing the dimensionality of data with neural networks. In *Science*.

Salter, D. A., Tamrakar, A., Behjat Siddiquie, M. R. A., Divakaran, A., Lande, B., & Mehri, D. (2015). The tower game dataset: A multimodal dataset for analyzing social interaction predicates. In *ACII*.

Schuller, B., et al. (2011). Avec 2011—the first international audio visual emotion challenge. In *ACII*.

Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2013). Affect analysis in natural human interactions using joint hidden conditional random fields. In *ICME*.

Sminchisescu, C., Kanaujia, A., & Metaxas, D. (2006). Learning joint top-down and bottom-up processes for 3d visual inference. In *CVPR*.

Srivastava, N., & Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. In *NIPS*.

Sun, X., Lichtenauer, J., Valstar, M. F., Nijholt, A., & Pantic., M. (2011). A multimodal database for mimicry analysis. In *ACII*.

Sutskever, I., & Hinton, G. E. (2007). Learning multilevel distributed representations for high-dimensional sequences. In *AISTATS*.

Sutskever, I., Hinton, G., & Taylor, G. (2008). The recurrent temporal restricted Boltzmann machine. In *NIPS*.

Taylor, G. W., et al. (2010). Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*.

Taylor, G. W., Hinton, G. E., & Roweis, S. T. (2011). Two distributed-state models for generating high-dimensional time series. *Journal of Machine Learning Research*, *12*, 1025–1068.

Wu, D. (2014). Deep dynamic neural networks for gesture segmentation and recognition. In *ECCV-W*.

Zanfir, M., Leordeanu, M., & Sminchisescu, C. (2013). The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*.

Zeiler, M. D., & Fergus, R. (2014). A multimodal database for mimicry analysis. In *ECCV*.

Zeiler, M. D., Taylor, G. W., Sigal, L., Matthews, I., & Fergus, R. (2011). Facial expression transfer with input–output temporal restricted Boltzmann machines. In *NIPS*.

Zhao, G., & Barnard, M. (2009). Lipreading with local spatiotemporal descriptors. *Transactions of Multimedia*, *11*, 1254–1265.